Published online 18 October 2010 *Nucleic Acids Research, 2011, Vol. 39, No. 4* ***1501–1509***
doi:10.1093/nar/gkq920

Identification of RNA recognition elements in the *Saccharomyces cerevisiae* transcriptome

Daniel P. Riordan[1,2,*], Daniel Herschlag[1] and Patrick O. Brown[1]

[1]Department of Biochemistry and [2]Department of Genetics, Stanford University School of Medicine, Stanford, California, USA

Received August 20, 2010; Revised September 15, 2010; Accepted September 25, 2010

ABSTRACT

Post-transcriptional regulation of gene expression, including mRNA localization, translation and decay, is ubiquitous yet still largely unexplored. How is the post-transcriptional regulatory program of each mRNA encoded in its sequence? Hundreds of specific RNA-binding proteins (RBPs) appear to play roles in mediating the post-transcriptional regulatory program, akin to the roles of specific DNA-binding proteins in transcription. As a step toward decoding the regulatory programs encoded in each mRNA, we focused on specific mRNA–protein interactions. We computationally analyzed the sequences of *Saccharomyces cerevisiae* mRNAs bound *in vivo* by 29 specific RBPs, identifying eight novel candidate motifs and confirming or extending six earlier reported recognition elements. Biochemical selections for RNA sequences selectively recognized by 12 yeast RBPs yielded novel motifs bound by Pin4, Nsr1, Hrb1, Gbp2, Sgn1 and Mrn1, and recovered the known recognition elements for Puf3, She2, Vts1 and Whi3. Most of the RNA elements we uncovered were associated with coherent mRNA expression changes and were significantly conserved in related yeasts, supporting their functional importance and suggesting that the corresponding RNA–protein interactions are evolutionarily conserved.

INTRODUCTION

How are precise patterns of gene expression reproducibly specified by molecular information encoded in genome sequences? A full understanding of the mechanisms and logic of this code requires systematic identification of individual regulatory elements encoded in the genome and characterization of the molecular interactions they impart and their regulatory consequences. The regulatory elements that specify the post-transcriptional regulation of each mRNA are still largely undiscovered. Hundreds of specific RNA-binding proteins (RBPs) now appear to be directly involved in regulating the post-transcriptional life of each mRNA (1,2). We therefore searched for the specific sequence elements recognized by a group of *Saccharomyces cerevisiae* RBPs.

Systematic studies of RNA–protein interactions have revealed that individual yeast RBPs typically associate with specific sets of mRNAs sharing related functional or cytotopic properties; many RNAs have been shown to interact with multiple RBPs, despite sparse experimental coverage of the universe of yeast RBPs (1,3–5). These observations suggest that combinatorial tagging of mRNAs via regulated interactions with diverse RBPs may be a general mechanism for specifying the distinct post-transcriptional fate of each mRNA in the cell (6).

To identify RNA elements recognized by specific yeast RBPs, we applied both bioinformatic and experimental approaches. We present a detailed description of our bioinformatic methodology, including additional analyses of the computationally predicted RNA motifs that we reported earlier (1). We also describe results using an *in vitro* selection approach to identify specific RNA sequences selectively bound by each of a dozen yeast RBPs. Characterization of these recognition motifs provides insight into how post-transcriptional regulatory information is encoded in the genome and facilitates analysis of the functional and evolutionary properties of these RNA elements.

MATERIALS AND METHODS

Bioinformatic motif analysis

Non-redundant sequence databases of putative 5′- and 3′-untranslated regions (UTRs) were generated and REFINE, MEME and FIRE motif prediction was performed as described (1). Full sequences are available, along with programs for running REFINE, in the Supplementary Data. Details of sequences and motif models used are in Supplementary Data S6. For both REFINE and FIRE, statistical significance of the

*To whom correspondence should be addressed. Tel: (650) 723 6719; Fax: (650) 725 7811; Email: driordan@stanford.edu

predicted motifs was assessed by randomly generating simulated target sets of similar size for each RBP and repeating the procedure 100 times on the simulated target data. We defined a test statistic as the negative $\log_{10}$ of the *P*-value for motif enrichment for REFINE; the reported motif *z*-score was used for FIRE motifs, and we compared the observed values of these test statistics to the distributions generated from the random simulations. Motifs were declared as significant if the observed test statistic was greater than three standard deviations above the mean, or if there was significant target-specific enrichment ($P < 10^{-4}$) of the motif in mRNA regions from which that motif was not originally predicted.

Preparation of cell extracts

TAP-tagged yeast strains (1–21) were grown in yeast-peptone-adenine-dextrose media to mid-log phase (OD600 = 0.6–0.9) and harvested as described earlier (3). Cell pellets were frozen in liquid N_2 and stored at −80°C. Pellets were cryogenically lysed in a Retsch MM301 ball mill and resuspended in Buffer B (100 mM Tris–HCl pH 8.0, 140 mM KCl, 1.8 mM $MgCl_2$, 0.1% NP-40 alternative, 0.2 mg/ml Heparin, 0.5 mM DTT, 1 mM PMSF, 0.5 μg/ml Leupeptin, 1.0 μg/ml Pepstatin, 10 U/ml SUPERasin). Lysates were cleared by centrifugation, resuspended at 10 mg/ml (protein), frozen in liquid N_2, then stored at −80°C until used.

RNA library generation

DNA oligonucleotide sequences used for SELEX libraries L1 and L2 were: T7-A1 = 5′-GCGTAATACGACTCACTATAGGGAGCATAGTTGCACGAGC-3′, R1 = 5′-CTATCATTCGGCAGACAGGCN(30)GCTCGTGCAACTATGCTCCC-3′, B1 = 5′-CTATCATTCGGCAGACAGGC-3′, A1 = 5′-GGGAGCATAGTTGCACGAGC-3′ and T7-A2 = 5′-GCGTAATACGACTCACTATAGGGAGACGATGGATGTCAAG-3′, R2 = 5′-CTGTGTCTTAGCAGCCGAACN(30)GTTCGGCTGCTAAGACACAG-3′, B2 = 5′-CTGTGTCTTAGCAGCCGAAC-3′ and A2 = 5′-GGGAGACGATGGATGTCAAG-3′. dsDNA was generated by a single primer extension reaction with 2 μM each of oligos T7-A1 and R1 for L1 (or T7-A2 and R2 for L2) under standard conditions with 5 U Platinum Taq in 50 μl (Invitrogen). Reaction products were used directly as template for standard T7 *in vitro* transcription reactions at 37°C for 2 h. RNA was isolated by Invitrogen Micro-to-Midi kit and quantitated by A260.

RNA *in vitro* selections

Invitrogen M-280 streptavidin-coated dynabeads (10 mg/ml) were prepared with biotinylated Rabbit IgG (EMD Biosciences) following standard procedures. IgG-beads were equilibrated with Buffer B and concentrated to 150 mg/ml. IVT library RNA (100 pmols) was heated at 70°C for 2 min then cooled on ice and added to 1.0 ml of thawed lysate along with 50 μl of IgG-beads. Binding reactions were carried out for 30 min at 25°C on a rotator. Beads were collected magnetically and washed three times on a rotator in 1.0 ml Buffer B for 10 min at 4°C, followed by three washes in 1.0 ml Buffer C (Buffer B with 10% glycerol and no heparin) for 10 min each at 4°C. Beads were then resuspended in a final volume of 200 μl before addition of 10 μl TEV protease (Invitrogen) and incubation at 18°C for 30 min. Beads were separated and the supernatant was collected and used for RNA isolation by Invitrogen Micro-to-Midi kit. RNA was eluted in 40 μl H_2O. Standard 20 μl thermoscript reverse transcription (RT) reactions were performed with 0.5 μM B1 primer at 60°C for 30 min using 10 μl of selected RNA as input. About 5 μl of RT reaction mixture was loaded into a 50 μl standard 20-cycle Platinum Taq PCR reaction with 2 μM each of oligos T7-A1 and B1. PCR product (10 μl) was used as a template in 50 μl T7 IVT reactions to produce new RNA for the subsequent selection. After four rounds of selection, the amount of library DNA in RT products was determined by qPCR using the Power SYBR Green kit (ABI) with 1.0 μl of RT product for each sample and 900 nM of primers A1 and B1. Absolute quantitation was performed using a standard curve made by 10-fold serial dilutions of known amounts of oligo R1 or R2.

Phylogenetic analysis of RNA motif sites

Genomic sequences from six related Saccharomyces species (*S. paradoxus, S. mikatae, S. kudriavzevii, S. bayanus, S. castellii, S. kluyveri*) were downloaded from the Saccharomyces Genome Database (http://www.yeastgenome.org). Separate multiple alignments were generated for the open reading frame (ORF) and 200 base upstream/downstream regions for each gene using CLUSTALW 1.82 (7). To estimate background conservation rates, permuted versions of each alignment were created by randomly shuffling the alignment columns while requiring that the resulting randomized alignment maintained the same *S. cerevisiae* sequence and pattern of gaps as the original alignment. Randomized ORF alignments were additionally constrained to require maintenance of the same encoded peptide sequence and the same codon usage as the original unshuffled alignments. Original and randomized alignments are available upon request, due to the large number of files.

RESULTS AND DISCUSSION

Computational inference of RNA recognition motifs from *in vivo* RBP target sequences

Computational identification of sequence features associated with particular biological properties has proven to be a powerful approach for the discovery of novel regulatory elements (8–12). Many algorithms have been developed to find significant motifs that occur frequently within a ‘target’ set of genomic sequences that share a functional characteristic, such as concordant expression profiles or protein interactions (8,10–12). A limitation of this kind of algorithm is the possibility of predicting non-specific motifs that occur at uniformly high levels in both target and non-target sequences, which makes it important to assess which predicted motifs are specifically enriched in the targets as compared with other genomic sequences. Alternative

algorithms avoid this problem by directly evaluating sets of potential motifs to find those with distributions that differ significantly between targets and non-targets in the genome, but use simpler regular expression models for motifs (9). To combine the strengths of these two approaches, we developed a new methodology called relative filtering by nucleotide enrichment (REFINE) to explicitly find target-specific motifs while accommodating position specific scoring matrix (PSSM)-based models. We used this approach to identify candidate RNA recognition elements within sets of 'target' RNAs bound *in vivo* by specific RBPs as determined by microarray analysis of RNAs enriched by immunopurification (IP) of each RBP (1).

REFINE first searches the specific RNAs identified as targets for segments that contain sequence patterns over-represented in the target set relative to the whole transcriptome, then uses existing tools to identify motifs in these segments. First, all possible nucleotide hexamers are evaluated for enrichment in the set of mRNAs bound by a specific RBP, compared to all other mRNAs (hyper-geometric $P < 10^{-3}$). Individual target sequences having at least three occurrences of significantly enriched hexamers are then chosen and subjected to additional steps. Segments comprising these enriched hexamers, along with three flanking residues on each side and intervening sequences of up to 12 bases that connect two adjacent hexamers are selected. Low-complexity regions with repetitive tri-nucleotide occurrences are filtered out using the program DUST (ftp.ncbi.nlm.nih.gov/pub/tatusov/dust). The resulting filtered target segments are used as input sequences for the MEME motif-finding algorithm (8). The motifs identified by MEME are then evaluated for specific enrichment in target RNAs and statistical significance is evaluated by stringent tests based on random simulations (details in 'Materials and Methods' section). Source code for REFINE is available in Supplementary Data.

Although REFINE is not guaranteed to completely avoid the problem of predicting potentially non-specific motifs, we nevertheless found it to be a useful step toward addressing this issue, as supported by the subsequent analyses. We also applied another motif-finding program, FIRE, to the same dataset (9). The overall concordance between the results of FIRE and REFINE (Supplementary Data S1) provided additional confidence in the significance and robustness of our results. All non-palindromic RNA motifs exhibited a strand bias, in that the reverse-complement motifs were not significantly enriched in the corresponding RBP target mRNAs (hyper-geometric $P > 0.01$) (Figure 1). This strand-specific enrichment is expected for regulatory elements that function as RNA, but not necessarily for DNA sequence motifs. Fourteen distinct RNA motifs, six of which (Puf3-1, Puf4-1, Puf5-1, Pub1-1, Nab2-1 and Nrd3-1) matched previously known RBP binding sites (Figure 1), passed strict criteria in this integrated analysis. The putative recognition motifs that we found for three of the RBPs (Pab1-1, Khd1-1 and Vts1-1) differed from the reported specificities of these RBPs (13–15), suggesting that these motifs may be false positives, perhaps

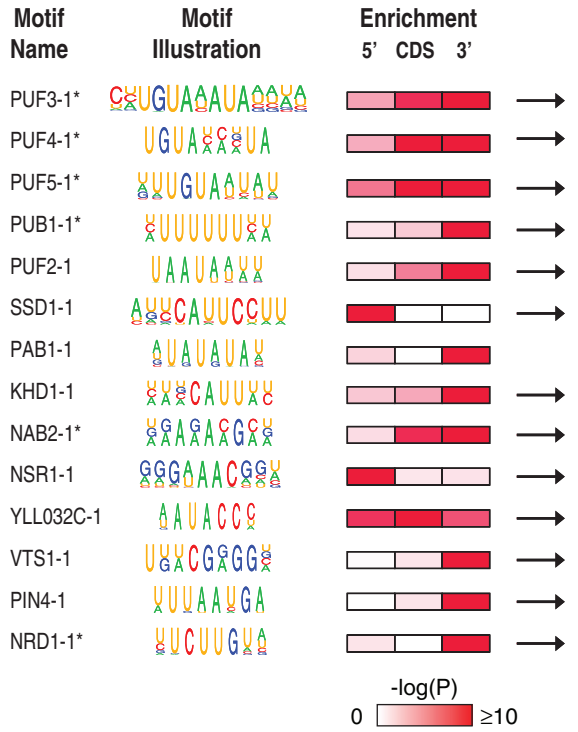

Figure 1. Computationally identified sequence motifs enriched in mRNAs bound by specific RBPs. RNA motifs identified from analysis of mRNA target sequences are displayed in decreasing order of significance based on *P*-values for genome-wide enrichment. A pictogram (http://genes.mit.edu/pictogram.html) represents the regular expression patterns defined for FIRE motifs or the preferred base composition of the position-specific scoring matrices used for REFINE motifs. For each motif, the $-\log_{10}$ *P*-value of the significance of genome-wide enrichment for motif sites in targets is shown in a red color scale for separate regions of its mRNA targets (5′ = 200 bases upstream of start codon, CDS = protein coding sequence, 3′ = 200 bases downstream of stop codon). Arrows are shown for motifs with a forward strand bias, i.e. the reverse complement of the motif is not significantly enriched in targets ($P > 0.01$). All relevant *P*-values were calculated based on the hyper-geometric distribution. Asterisks denote motifs that correspond to previously reported binding sites for the associated RBP. Exact data values and supporting details are presented in Supplementary Data S1.

representing sequences recognized by other factors with similar sets of target genes. The remaining motifs are strong candidates for specific RNA elements bound by Puf2, Ssd1, Nsr1, YLL032C and Pin4, respectively. Several of the motifs predicted by REFINE (including Puf5-1, Puf2-1, Nsr1-1, YLL032C-1, Vts1-1, Pin4-1 and Nrd1-1) were not identified by standard MEME analysis of the same original input target sequences, suggesting that analyses using MEME alone may be unlikely to recover some of these elements (Supplementary Data S1).

Our analysis assumed that the 200 bases upstream of the start codon and 200 bases downstream of the stop codon, respectively, defined the 5′- and 3′-UTRs of the mRNAs. When precise experimental annotations of the UTR boundaries from mRNA-seq data later became available (16), we tested whether RNA motifs that we had identified in the 200 base putative UTR regions tended to reside within the actual annotated UTRs more often than

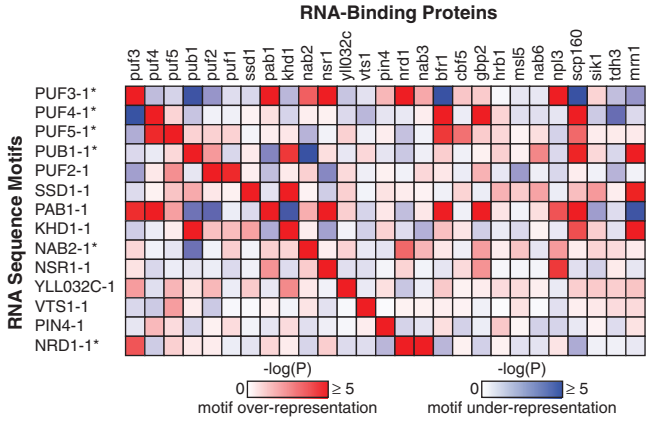

Figure 2. Enrichment of RNA motifs in diverse RBP target sets. A heatmap illustrates the degree to which each of the identified RNA sequence motifs from Figure 1 (rows) is significantly over-represented or under-represented in the sequences of the set of target mRNAs bound by different RBPs (columns). Cells are either shaded in red to indicate the $-\log_{10}$ *P*-value of the significance of over-representation of motifs, or likewise shaded in blue for under-represented motifs. The red squares along the diagonal reflect the fact that each RNA motif was originally defined based on its strong over-representation within the target sequences of its cognate RBP. All *P*-values were calculated based on the hyper-geometric distribution. Asterisks denote motifs that correspond to previously reported binding sites for the associated RBP. Exact data values and supporting details are presented in Supplementary Data S2.

expected by chance. The motifs for Puf3, Puf4, Puf5, Puf2, Ssd1, Pab1 and Pin4 all demonstrated a significant propensity to occur within the annotated UTRs (Supplementary Data S1). In contrast, the motif sites for Nrd1/Nab3 and Nab2 were significantly more likely to be located downstream of the annotated 3′-UTRs, suggesting these nuclear RBPs may function by recognizing sequences near the 3′-end of nascent RNAs prior to cleavage and poly-adenylation. As the Nrd1/Nab3 complex controls transcription termination of nascent RNA substrates (17), the tendency for its motifs to occur outside of mature transcripts appears to reflect its established biological role.

Our analysis also revealed some motifs that were repeatedly enriched in the RNA targets of multiple RBPs (Figure 2). This observation may be a clue to interconnections in the regulatory network underlying post-transcriptional regulation (1). As a practical matter, however, the potential for spurious motif enrichment due to functionally overlapping regulons can make it difficult to infer which element specifically interacts with a RBP of interest. The sequence biases of protein-coding regions and the inherent difficulty of predicting RNA structures from primary sequences pose additional challenges to computational methods and highlight the value of direct experimental tests of RBP sequence specificity (18).

In vitro selection of RNA recognition elements

We developed an efficient SELEX (systematic evolution of ligands by exponential enrichment) (19,20) protocol for selecting RNA ligands that specifically bind to individual RBPs. The approach takes advantage of the yeast genome-wide collection of TAP-tagged strains (21). Binding reactions were performed by adding *in vitro*-transcribed RNA pools, consisting of 30 randomized bases flanked by two 20 base constant regions, to a cell lysate containing the TAP-tagged RBP of interest. Each reaction contained $\sim 6 \times 10^{13}$ molecules of library RNA, which theoretically represents ~600-fold coverage of all 20-mers in the randomized pool. In each round of selection, the RBP and its associated RNAs were selectively isolated, the recovered RNAs were reverse transcribed and the products were amplified by PCR using primers specific to the flanking constant sequences. Four cycles of selection were performed for each RBP. At each cycle, the fraction of input library selectively bound by the RBP was monitored by qPCR. All selections were performed using two distinct libraries (L1 and L2) with different constant sequences to facilitate the detection of library-specific features that could have contributed to the selection. For 10 of the 12 RBPs tested (excepting Bfr1 and Khd1), serial enrichment yielded RNA pools with apparent affinity significantly above the background level in the unselected pool for both libraries (Supplementary Data S3). We inferred RNA recognition elements by manually analyzing the sequences of individual molecules enriched in each selection, then checking for enrichment of the inferred motifs in the empirically derived *in vivo* RNA targets of the corresponding RBP (Figure 3).

We first evaluated our SELEX approach by inspecting the *in vitro* selected sequences obtained for four RBPs in our study with recognition motifs described earlier: Puf3, Vts1, Whi3 and She2. The sequences selected by Puf3-binding from both libraries yielded a consensus motif that closely matched the previously identified Puf3 binding site, UGUAHAUA (H = A/C/U) (27/28 clones from the L1 pool, $P = 8.0 \times 10^{-61}$ and 9/9 of L2 clones, $P = 1 \times 10^{-27}$) (3,22). Vts1 has been reported to bind to 'Smaug Recognition Element' (SRE) motifs, consisting of a short hairpin with a $\geq$4-bp stem and a tetra- or penta-loop with sequence CNGGN(0–1) (13,23). Indeed, we found sequences conforming to the canonical SRE model enriched in the Vts1-selected clones from both libraries (23/28 L1, $P = 4.8 \times 10^{-34}$ and 2/6 L2, $P = 0.0073$). Moreover, the sequences selected by Vts1 *in vitro* displayed a strong bias for an oriented G:C base pair at the top of the hairpin stem (21/23 L1, $P = 6.9 \times 10^{-10}$), consistent with bioinformatic evidence that this feature is involved in Vts1 recognition (24). In fact, while canonical SREs were significantly enriched in the empirical *in vivo* targets of Vts1 ($P < 10^{-8}$), SREs with the G:C base pair were even more significantly enriched ($P < 10^{-21}$). Sequences selected from both libraries by binding to Whi3 yielded UGCAU as a consensus motif (16/19 L1 clones, $P = 4.4 \times 10^{-23}$ and 4/12 L2 clones, $P = 2.0 \times 10^{-4}$), extending the earlier reported GCAU recognition motif for Whi3 (25). Indeed, UGCAU sites were more significantly enriched in the 3′-UTRs of mRNAs bound by Whi3 ($P = 8.1 \times 10^{-14}$) than were GCAU sites ($P = 1.1 \times 10^{-6}$), suggesting that the additional U on the 5′-edge of the motif contributes to

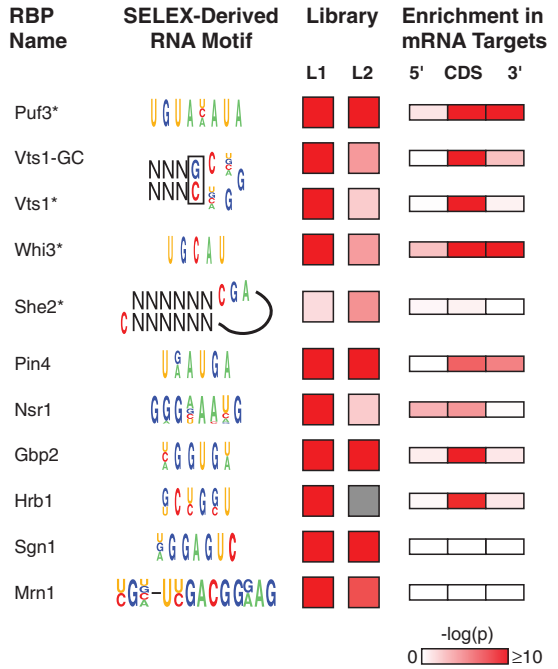

Figure 3. RNA recognition elements determined by *in vitro* selection. RNA motifs determined by analysis of SELEX clone sequences are depicted for each RBP. The $-\log_{10}$ *P*-values of the significance of motif enrichment in clones from the two distinct SELEX libraries ('L1' and 'L2') are represented in a red color scale. For each motif, the $-\log_{10}$ *P*-value of the significance of genome-wide enrichment for motif sites in segments of its mRNA targets bound *in vivo* is also color-coded. Motifs are listed in the order they are discussed in the text. The gray box indicates data are not available because all L2 clones inadvertently contain an Hrb1 motif site in their 3′ constant region. Asterisks denote motifs that correspond to previously reported binding sites for the associated RBP. For exact data values and details see Supplementary Data S3.

in vivo target recognition (25). Sequences selected by the She2 protein from L2 were enriched (10/22, $P = 4.8 \times 10^{-5}$) for the earlier reported She2 'zipcode' motif, consisting of a loop–stem–loop structure with a CGA sequence in one loop precisely positioned across from a single C residue in the other loop (Figure 3) (26,27). Despite the high background frequency of this motif in the L1 library (58% of unselected sequences), we found significant enrichment of a variant of the zipcode motif, with dual CGA sites in both loops (8/28, $P = 4.6 \times 10^{-8}$). The results for known RBP recognition elements show that this procedure can faithfully identify specific RBP recognition elements of diverse sequences and structures.

Our bioinformatic analyses predicted novel RNA recognition element motifs for the Pin4 and Nsr1 proteins, which we could compare to the results of the corresponding SELEX experiments. The sequences selectively bound by Pin4 from both libraries were enriched for URAUGA sites (R = A/G) (8/17 L1, $P = 1.6 \times 10^{-11}$ and 9/19 L2, $P = 7.6 \times 10^{-13}$) similar to the motif predicted from analysis of its *in vivo* targets (Figure 1), strong corroboration for this motif as a Pin4 recognition element. Likewise, a sequence with the consensus GGGNAANG matching the computationally predicted Nsr1 motif (Figure 1) was enriched in both libraries (38/58 L1, $P = 1.3 \times 10^{-88}$ and 2/20 L2, $P = 0.009$). The significant divergence of this motif from the sequences recognized by its mammalian homolog Nucleolin may account for the inability of mammalian Nucleolin to rescue yeast *nsr1* mutants (28). How the evolutionary 'rewiring' of this critical RBP occurred is an interesting problem for future investigations.

The binding specificities of two paralogous yeast RBPs, Gbp2 and Hrb1, present another potential example of evolutionary rewiring. Our computational analysis did not turn up a convincing candidate recognition element for either of these serine–arginine rich proteins, both of which are involved in mRNA export (29). From the sequences selected by Gbp2 binding *in vitro*, we detected a novel motif, with consensus HGGUGW (H = A/C/U, W = A/U), significantly enriched in both libraries (13/24 L1, $P = 8.9 \times 10^{-13}$ and 15/20 L2, $P = 2.3 \times 10^{-15}$). Moreover, sequences matching the HGGUGW motif were significantly more frequent in the coding regions of mRNAs bound by Gbp2 *in vivo* than in coding regions from the transcriptome at large (0.23% versus 0.16% of hexamers, Poisson $P = 4.1 \times 10^{-47}$). A motif fitting the consensus KCYGSU (K = G/U, Y = C/U, S = C/G) was enriched in the sequences selected by Hrb1 binding *in vitro* (16/20 L1, $P = 2.7 \times 10^{-16}$), and also in the coding regions of *in vivo* target mRNAs (0.33% versus 0.16% of hexamers, Poisson $P = 1.5 \times 10^{-9}$), suggesting that this motif represents a *bona fide* recognition element for Hrb1. The differences in the recognition motifs and the target RNAs for Gbp2 and Hrb1 suggest that, following their ancestral gene duplication, these paralogous RBPs diverged in specificity and biological roles. The overlapping region of their recognition motifs (KCYGSU and HGGUGW, respectively) may be a vestige of the binding of their common progenitor.

Sgn1 and Mrn1 are two RBPs for which our computational analysis did not identify high-confidence candidate recognition motifs. The motif DGGAGUC (D = A/G/U) was greatly enriched in both Sgn1 SELEX libraries (23/27 L1, $P = 1.3 \times 10^{-49}$, and 16/21 L2, $P = 5.6 \times 10^{-32}$). A motif with consensus YGN(0–4)UYGACGGRAG (Y = C/U, R = A/G) was enriched in both Mrn1 SELEX libraries (3/16 L1, $P = 1.9 \times 10^{-10}$, and 2/6 L2, $P = 5.4 \times 10^{-8}$). Neither of these *in vitro* selected motifs, however, was detectably enriched in the reported *in vivo* targets of the cognate RBPs. These motifs could bind the RBPs in a manner that differs from the interactions with *in vivo* targets; further experiments will be needed to test this and other possibilities.

Phylogenetic conservation of RNA interaction networks

Our bioinformatic analysis and SELEX experiments identified RNA motifs that are likely to mediate specific regulatory interactions between RBPs and their associated mRNA targets. If these putative regulatory elements are functionally important, then natural selection may have constrained the evolution of these motif sites in

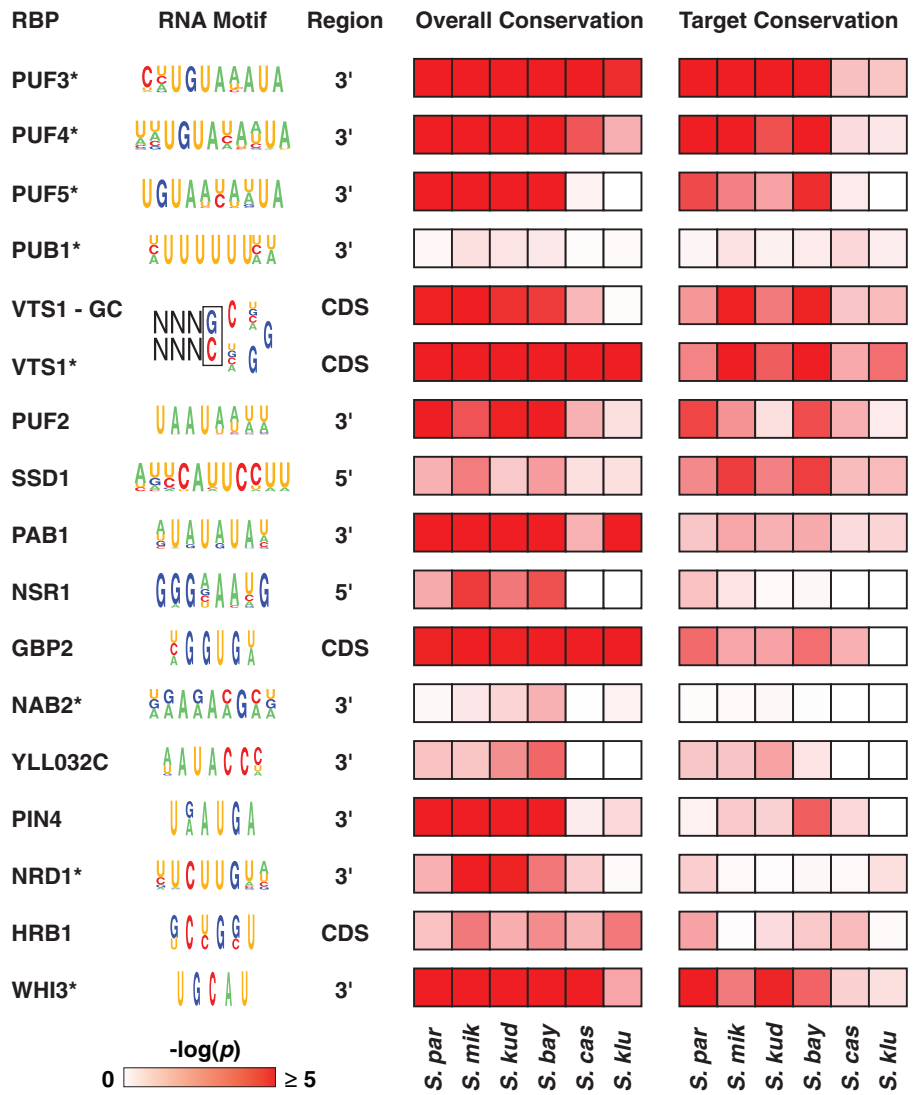

Figure 4. Conservation of RNA elements in *Saccharomyces*. Phylogenetic conservation rates of all sites for each RNA motif were calculated between *S. cerevisiae* and each of six related *Saccharomyces* species (par, *paradoxus*; mik, *mikatae*; kud, *kudriavzevii*; bay, *bayanus*; cas, *castellii*; klu, *kluyveri*) based on multiple alignments of the indicated genomic regions. For 'Overall Conservation', each cell is shaded according to the $-\log_{10}$ *P*-value measuring if the observed conservation rate of motif sites in that species is significantly greater than expected by chance based on randomized alignments. For 'target conservation', the cell for each species is shaded to depict the $-\log_{10}$ *P*-value measuring if the conservation rate of motif sites present within sequences of target mRNAs bound by the cognate RBP is significantly greater than the conservation rate of motif sites from all other transcripts. All *P*-values were calculated based on the hyper-geometric distribution. Asterisks denote motifs that correspond to previously reported binding sites for the associated RBP. For exact data values and details see Supplementary Data S4.

order to preserve their functions in closely related yeasts. Evidence for such purifying selection has been repeatedly observed and interpreted as evidence for the functional importance of proposed transcription factor binding sites (30). We thus reasoned that phylogenetic analysis could similarly provide an independent test to evaluate the RNA motifs we identified.

To examine the evolutionary properties of the putative RBP recognition elements, we generated separate multiple sequence alignments for each region (i.e. ORF, 200 bases upstream of start codon or 200 bases downstream of stop codon) of every gene using sequences from *S. cerevisiae* and six other *Saccharomyces* species (*S. paradoxus, S. mikatae, S. kudriavzevii, S. bayanus, S. castellii* and *S. kluyveri*) (31,32). To control for differences in average conservation rates between different regions of genes, the conservation of each motif was examined separately for each region. For each species, the orthologous sequences that aligned to the site of a candidate motif in *S. cerevisiae* were evaluated using the same motif model. The overall conservation rate for each species was defined as the fraction of all sequences orthologous to a motif site in *S. cerevisiae* that also satisfied the motif model. This strict alignment-based conservation rate demands that both conformity to the motif model and position in the aligned sequence be conserved, excluding instances in which the presence of a motif in a transcript, but not its location, is conserved.

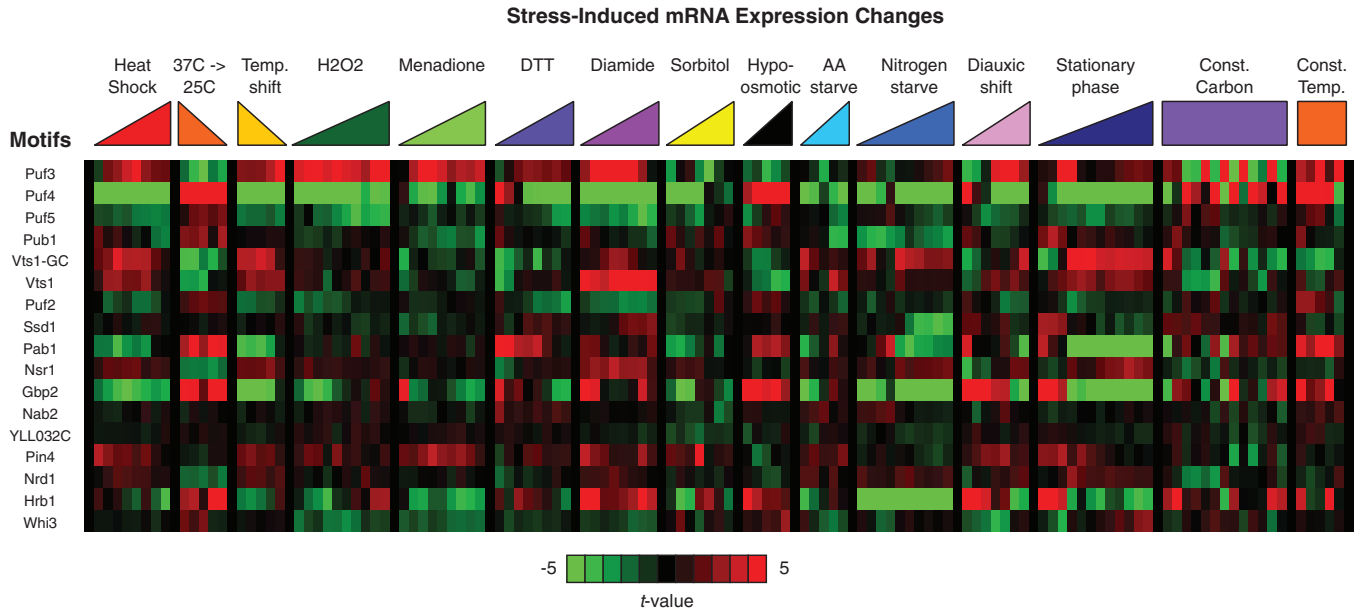

Figure 5. mRNA expression changes associated with RNA motifs. A heatmap illustrates the degree to which the relative expression levels of mRNAs containing each of the identified RNA sequence motifs (rows) changed under each of the environmental stress conditions shown (columns). For each motif and each stress condition we calculated the *t*-value measuring how much the average expression change of mRNAs with motif sites deviated from its expected value by chance. Relative increases in average mRNA expression levels are colored in red, and relative decreases are colored in green. For exact data values and supporting details see Supplementary Data S5.

Using these criteria, we found significant evidence of conservation for almost all of the RNA motifs we tested—including the new SELEX-derived recognition elements for Gbp2, Hrb1, Nsr1 and Pin4, and the novel predicted motifs for Ssd1 and Puf2—in at least one related species ($P < 0.01$) (Figure 4). Statistical significance of overall conservation rates was established based on a null model using permuted versions of each alignment file that were carefully constrained to preserve the identical per-nucleotide conservation rates from the original alignments ('Materials and Methods' section). As an additional analysis, we also asked if the conservation rate for each RNA motif was significantly greater for the collection of sites present in the mRNA targets of its cognate RBP than for sites present in all other transcripts. We found significant evidence of this type of 'target-specific' preferential conservation for many putative RBP recognition elements—including the motifs for Puf3, Puf4, Puf5, Vts1, Puf2, Ssd1, Gbp2, Pin4 and Whi3 ($P < 0.01$) (Figure 4). Together, this evidence for purifying selection acting to preserve functional recognition elements during evolution suggests that the corresponding post-transcriptional regulatory networks specified by these elements have been conserved among *Saccharomyces* species as well.

RNA motifs are associated with mRNA expression patterns

Our bioinformatic and SELEX results support the idea that the RNA motifs we identified play a role in encoding an extensive network of RNA–protein interactions involving virtually all mRNAs in the transcriptome. The significant evolutionary conservation of the RNA motifs we found suggests that the interactions specified by these elements are not merely decorative, but likely confer regulatory functions that contribute to organismal fitness. However, the regulatory programs that we hypothesize may be mediated by these RNA–protein interactions remain largely uncharacterized.

To investigate the potential roles of the identified RNA motifs in condition-specific regulation of mRNA abundance, we integrated our motif results with a large-scale compendium of stress-induced gene expression programs (33). For each environmental stress condition and each RNA motif, we calculated a *t*-statistic value (34) as a measure of the direction and significance of coherent changes in expression of the mRNAs that share a putative RBP recognition element (Figure 5). This analysis revealed numerous conditions under which sets of genes defined by a common RNA motif showed coherent stress-induced changes in mRNA abundance [263 motif-condition pairs for which (*t*-value) > 4, FDR $< 10^{-3}$].

The overall links observed here between putative RBP recognition elements and expression patterns strengthen the evidence that these elements have a regulatory role and suggest the possibility that RNA–protein interactions mediated by these sites could contribute directly to these condition-specific alterations in mRNA levels (presumably by condition-specific regulation of mRNA decay). Future experiments to dissect the influence of these RBPs on gene expression, and the molecular mechanisms through which they act, will be aided by the ability to focus on relevant growth conditions and to target specific recognition sites for mutagenesis.

SUMMARY AND PERSPECTIVE

We identified RBP recognition elements in the yeast transcriptome using an integrated computational and experimental strategy. We developed a new algorithm, REFINE, which we applied in conjunction with other programs to predict RNA motifs from sequences of mRNAs that were selectively bound by specific RBPs. Our computational results recovered the known binding specificities for Puf3, Puf4, Puf5, Pub1, Nab2 and Nrd1/Nab3, as well as strong candidate motifs for Puf2/Puf1, Ssd1, Nsr1, Mrn1 and Pin4. We also performed *in vitro* selections with a diverse set of 12 RBPs to biochemically characterize their specificities. Our SELEX results agreed with the previously reported motif models for Puf3, Vts1, Whi3 and She2. The SELEX data also revealed novel RNA motifs for Pin4, Nsr1, Gbp2 and Hrb1, each of which were enriched in the *in vivo* mRNA targets of the cognate RBPs.

The behavior of transcripts containing RBP recognition motifs in existing RNA expression datasets yielded clues to the potential regulatory functions of these elements and their cognate RBPs. Most of the RNA motifs we identified were significantly conserved in related *Saccharomyces* species, particularly when they occurred in orthologs of the experimentally identified RBP targets. This conservation provides independent evidence for the functional importance of the RNA elements that we identified and suggests that the corresponding regulatory interactions are preserved in related species. Identification of the respective RBP recognition elements also revealed apparent divergence in the specificities of the paralogous RBPs Gbp2 and Hrb1, and the orthologous RBPs Nsr1 and mammalian Nucleolin.

This study also highlights some of the practical issues relevant to the identification of RBP recognition elements. For example, structured recognition elements (Vts1 and She2), and recognition elements located in coding sequences (Gbp2 and Hrb1) were missed by our computational analysis, which was directed at primary sequence features in UTRs. The fact that even the well-studied She2 zipcode elements are not statistically significantly enriched in the bud-localized She2 target RNAs clearly highlights the difficulty of inferring RNA structure from primary sequence data alone. Although we overcame some of these issues by using SELEX, *in vitro* selection experiments also have known limitations (35). One drawback is the potential selection of RNA or DNA features that are enriched during the procedure for unintended reasons. We saw evidence for selection of a biotin-binding aptamer in the L2 Hrb1 selection, which was presumably selected for binding to the biotinylated IgG that we used to capture the TAP-tagged RBP (36). From RNA folding predictions of *in vitro* selected sequences (37), we also detected a propensity for structured regions flanking the SELEX motifs for Puf3, Whi3, Gbp2 and Hrb1, a feature not thought to be associated with corresponding motifs in the *in vivo* targets (Supplementary Data S3). These examples underscore the importance of combining our *in vitro* binding results with information about the *in vivo* biological target RNAs. Newer methods for large-scale empirical identification of the binding sites and specificities of RBPs have recently been reported which, while costlier, may circumvent some of the limits of the methods used in this study (38–40).

Post-transcriptional regulation of RNA mediated by RBPs is a common mode of control in the global gene expression program. Further characterization of RNA recognition elements, and the functional consequences of the interactions they specify, will help elucidate how distinct post-transcriptional regulatory fates are programmed in the sequence of each mRNA in the cell.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank Nina Tsvetanova and other members of the Brown lab for helpful advice and comments on the manuscript. P.O.B is an investigator for the Howard Hughes Medical Institute.

FUNDING

Howard Hughes Medical Institute; National Cancer Institute (R01 CA77097–08 to P.O.B.). Stanford Graduate Fellowship (to D.P.R.); Stanford Genome Training Program, National Human Genome Research Institute (Grant Number T32HG00044). Funding for open access charge: Howard Hughes Medical Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Hogan,D.J., Riordan,D.P., Gerber,A.P., Herschlag,D. and Brown,P.O. (2008) Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol*., **6**, e255.
2. Keene,J.D. (2007) RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet*., **8**, 533–543.
3. Gerber,A.P., Herschlag,D. and Brown,P.O. (2004) Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol*., **2**, E79.
4. Hieronymus,H. and Silver,P.A. (2003) Genome-wide analysis of RNA–protein interactions illustrates specificity of the mRNA export machinery. *Nat. Genet*., **33**, 155–161.
5. Shepard,K.A., Gerber,A.P., Jambhekar,A., Takizawa,P.A., Brown,P.O., Herschlag,D., DeRisi,J.L. and Vale,R.D. (2003) Widespread cytoplasmic mRNA transport in yeast: identification of 22 bud-localized transcripts using DNA microarray analysis. *Proc. Natl Acad. Sci. USA*, **100**, 11429–11434.
6. Keene,J.D. and Tenenbaum,S.A. (2002) Eukaryotic mRNPs may represent posttranscriptional operons. *Mol. Cell*, **9**, 1161–1167.
7. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*., **31**, 3497–3500.
8. Bailey,T.L. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn*., **21**, 51–80.
9. Elemento,O., Slonim,N. and Tavazoie,S. (2007) A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell*, **28**, 337–350.

10. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
11. Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.
12. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
13. Aviv,T., Lin,Z., Lau,S., Rendl,L.M., Sicheri,F. and Smibert,C.A. (2003) The RNA-binding SAM domain of Smaug defines a new family of post-transcriptional regulators. *Nat. Struct. Biol.*, **10**, 614–621.
14. Hasegawa,Y., Irie,K. and Gerber,A.P. (2008) Distinct roles for Khd1p in the localization and expression of bud-localized mRNAs in yeast. *RNA*, **14**, 2333–2347.
15. Mangus,D.A., Evans,M.C. and Jacobson,A. (2003) Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome Biol.*, **4**, 223.
16. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.
17. Steinmetz,E.J., Conrad,N.K., Brow,D.A. and Corden,J.L. (2001) RNA-binding protein Nrd1 directs poly(A)-independent 3′-end formation of RNA polymerase II transcripts. *Nature*, **413**, 327–331.
18. Mathews,D.H. and Turner,D.H. (2006) Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.*, **16**, 270–278.
19. Ellington,A.D. and Szostak,J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818–822.
20. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
21. Ghaemmaghami,S., Huh,W.K., Bower,K., Howson,R.W., Belle,A., Dephoure,N., O'Shea,E.K. and Weissman,J.S. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.
22. Olivas,W. and Parker,R. (2000) The Puf3 protein is a transcript-specific regulator of mRNA degradation in yeast. *EMBO J.*, **19**, 6602–6611.
23. Aviv,T., Lin,Z., Ben-Ari,G., Smibert,C.A. and Sicheri,F. (2006) Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1p. *Nat. Struct. Mol. Biol.*, **13**, 168–176.
24. Foat,B.C. and Stormo,G.D. (2009) Discovering structural cis-regulatory elements by modeling the behaviors of mRNAs. *Mol. Syst. Biol.*, **5**, 268.
25. Colomina,N., Ferrezuelo,F., Wang,H., Aldea,M. and Gari,E. (2008) Whi3, a developmental regulator of budding yeast, binds a large set of mRNAs functionally related to the endoplasmic reticulum. *J. Biol. Chem.*, **283**, 28670–28679.
26. Jambhekar,A., McDermott,K., Sorber,K., Shepard,K.A., Vale,R.D., Takizawa,P.A. and DeRisi,J.L. (2005) Unbiased selection of localization elements reveals cis-acting determinants of mRNA bud localization in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **102**, 18005–18010.
27. Olivier,C., Poirier,G., Gendron,P., Boisgontier,A., Major,F. and Chartrand,P. (2005) Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud. *Mol. Cell Biol.*, **25**, 4752–4766.
28. Ginisty,H., Sicard,H., Roger,B. and Bouvet,P. (1999) Structure and functions of nucleolin. *J. Cell Sci.*, **112(Pt 6)**, 761–772.
29. Hurt,E., Luo,M.J., Rother,S., Reed,R. and Strasser,K. (2004) Cotranscriptional recruitment of the serine-arginine-rich (SR)-like proteins Gbp2 and Hrb1 to nascent mRNA via the TREX complex. *Proc. Natl Acad. Sci. USA*, **101**, 1858–1862.
30. Pennacchio,L.A. and Rubin,E.M. (2001) Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.*, **2**, 100–109.
31. Cliften,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
32. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
33. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell.*, **11**, 4241–4257.
34. Boorsma,A., Foat,B.C., Vis,D., Klis,F. and Bussemaker,H.J. (2005) T-profiler: scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Res.*, **33**, W592–W595.
35. Johns,G.C. and Joyce,G.F. (2005) The promise and peril of continuous in vitro evolution. *J. Mol. Evol.*, **61**, 253–263.
36. Wilson,C., Nix,J. and Szostak,J. (1998) Functional requirements for specific ligand recognition by a biotin-binding RNA pseudoknot. *Biochemistry*, **37**, 14410–14419.
37. Hofacker,I.L. (2004) RNA secondary structure analysis using the Vienna RNA package. *Curr. Protoc. Bioinformatics*, Chapter 12, Unit 12, 12.
38. Granneman,S., Kudla,G., Petfalski,E. and Tollervey,D. (2009) Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc. Natl Acad. Sci. USA*, **106**, 9613–9618.
39. Licatalosi,D.D., Mele,A., Fak,J.J., Ule,J., Kayikci,M., Chi,S.W., Clark,T.A., Schweitzer,A.C., Blume,J.E., Wang,X. *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464–469.
40. Ray,D., Kazan,H., Chan,E.T., Pena Castillo,L., Chaudhry,S., Talukder,S., Blencowe,B.J., Morris,Q. and Hughes,T.R. (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.*, **27**, 667–670.